ABSTRACT

A method, system and computer readable medium for retrieving relevant data in large collections of documents is disclosed. The method, system and computer readable medium of the present invention includes retrieving a document to be indexed, generating a document extract from the document, wherein the document extract comprises a portion of the document, and decomposing the document extract into tokens. The tokens are then stored in a search index, wherein a search engine accesses the search index to retrieve information satisfying a search query.

Through aspects of the method, system and computer readable medium of the present invention, the quality of the search result is improved because the retrieved documents are more relevant in view of the semantic concept or notion represented by the search query. Moreover the storage requirements are reduced, while expediting the processing time for conducting a search.